

1

LOOKUP ENGINE

TECHNICAL FIELD

- 5 The present invention relates to a look up engine for use in computer systems. In particular, but not exclusively, it relates to look-up engine for use in routing tables, flow tables and access control lists.

10 BACKGROUND OF THE INVENTION

- 15 One area in which look up tables are extensively used are in routing tables for use by a router. A router is a switching device which receives a packet, and based on destination information contained within the data packet, routes the packet to its destination.

- 20 Each packet contains a header field and data field. The header field contains control information associated with the routing of the packet including source and destination information. On receiving a packet, a router identifies the key in the header field. The key contains the information that is used to look up the route for the received packet.

- 25 The look up table includes a plurality of entries having a route destination associated with a "key". After a key for a packet has been determined, the router performs the look-up in the look up table for the matching entry and hence the destination associated with the key and routes the packet accordingly. A given
30 key may typically match a large number of routes in the look up table.

- 35 Traditional routing processes using a conventional look up table are very time consuming. One known method to speed up this look up process is to cache the most recent or often performed matches.

Furthermore it is difficult to update conventional look up tables to change routing information.

One solution to this is to provide a look up table in which the entries are stored in a special format, known as a "trie". A trie is a multi-way tree structure used for organising data to optimise lookup performance. The data is organized as a set of linked nodes, in a tree structure. Each trie node contains a power-of-two number of entries. Each entry is either empty or contains the lookup result. If the entry is empty, it will point to another trie node and the look up process is repeated. If the entry contains the look up value, this value is returned and the look up process is effectively terminated.

A particular form of such a trie is a level-compressed trie (LC-trie) data structure also known as a "Patricia" tree (Practical Algorithm to Retrieve Information Coded In Alphanumeric).

A traditional trie uses every part (bit or characters) of the key. in turn, to determine which subtree to select. However, a Patricia tree nominates (by storing its position in the node) which element of the key will next be used to determine the branching. This removes the need for any nodes with just one descendent and consequently the Patricia tree utilises less memory than that required by a traditional trie. However, Patricia trees are fairly expensive to generate, so a table which utilises such a format is best used in applications for which lookup speed is more important than update speed. However, with increasing complexity of routers and hence the increased size of such look tables, it has become increasingly important to increase the speed of look up and the accuracy of lookup.

35 SUMMARY OF THE INVENTION

The object of the present invention is to provide a look up engine and look up process which provides fast and accurate look up.

This is achieved in accordance with a first aspect of the present invention by providing a look up table comprising a plurality of parallel look up state machine which can provide concurrent look ups. Each look up state machine accesses storage means,

preferably comprising a plurality of parallel, independent memory banks, in which the look up table may be constructed on the basis of a trie, more preferably a Patricia tree structure. Such a look up table provides increased performance by doing multiple parallel lookups to multiple memory banks in parallel. The returned value may be a final value or reference to another table.

10 The object of the invention is also achieved in accordance with a second aspect of the present invention by providing each trie entry with a skip value field. This enables the ability to avoid false hits, avoiding a memory access to check if a table hit is real. Conventional tries return false hits. During the lookup process, the skip value field is compared to the skipped key bits, and a lookup failure is signalled if they do not match. In the traditional implementation of LC-tries, skip values are not stored in the trie entries, which gives rise to false hits in the table. The possibility of false hits means that hits have to be confirmed by performing an additional memory reference to the full table. The provision of a skip value field for each entry eliminates the need for this extra memory reference, at the expense of somewhat larger entries. The look up engine in accordance with the first aspect may incorporate the feature of the second aspect. If the feature of the second aspect is not incorporated, then it can be appreciated that the false hits may be returned but the memory required for the look up table or tables would be reduced. Further, it can be appreciated that further processing would be required to detect such false hits.

35 Key lengths, for example, can be up to 128 bits and values can be up to 41 bits. The table lookup engine has some internal memory for table storage, and it can also use memory external to the table lookup engine block.

40 The object of the invention is also achieved in accordance with a third aspect of the present invention by providing a table lookup engine which deals with longest prefix matching by pre-processing the entries to split overlapping ranges. The conventional method is to maintain a "history stack" in the trie hardware for this. In pre-processing the entries in this way, the hardware is simplified.

In the event of multiple tables which may be used for different protocols, then these could be stored as separate tables and which table to be search is chosen by the value of the input key. Alternatively, the
 5 tables may be combined in the same tree so the first look up (and therefore the first bits of the input key value) is which way to branch to get at the appropriate sub-table.

10 Multiple logical tables can be supported simultaneously by prep ending the keys with a table selector.

The table lookup engine according to the present invention is capable of returning the number of bits
 15 that did match in the case of a table miss.

Parallel lookups can be further accelerated by pre-processing the tables, such that lookups that require more memory accesses have their entries preferentially
 20 placed in fast, on-chip RAM.

Further, in accordance with a preferred embodiment, the lookup table or tables is constructed in software giving a high degree of flexibility, for example, the
 25 length of the key value can be fixed or of variable length, the tree depth is programmable and the size of the tree and performance can be optimised. It is simply to design the look up with or without the facility of minimising false hits. Of course, it can be appreciated
 30 that a table which has false hits would be smaller in size, but would require further processing of the result to detect false hits. The software utilised by the present invention pre-processes the data into the trie structure which enables different performance
 35 trade-offs and types of lookups/return values possible with the same simple hardware.

BRIEF DESCRIPTION OF DRAWINGS

40 Figure 1 is a schematic block diagram of the LC-trie data structure of the look up table according to an embodiment of the present invention; and

45 Figure 2 is a schematic block diagram of the table look up engine according to the embodiment of the present invention.

50

1007403-021400

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

With reference to figure 1, the trie structure of a look up table according to an embodiment of the present invention will be described. The look up table comprises a plurality of entries 110a-110d, 120a-120h. Each entry comprises a look up value and an associated key value. The entries are arranged within the look up table in a plurality of hierarchal nodes, for example a first level 110 and a second level 120. Although only two levels are illustrated here, it can be appreciated that any number of levels may be supported.

A key 100 is input into the look up table. A predetermined number of the leading bits of the input key 100 are used to index into the first level 110 of the hierarchy of nodes. This is done by adding the value of these bits to the base address of the node. In the example shown in figure 1, the leading bits 101 of the input key 100 point to an entry 110b of the first level of nodes 110. The entry 110b contains a skip count and a skip value. The skip count causes the look up process to skip a predetermined number of bits 102 in the input key 100. The skip value indicates the number of bits 103 to be used to index into the next level 120 of nodes. As in the previous level the look up is carried out by adding the value of these bits 103 to the base address of the node 120. This points to a particular entry 120f. This entry 120f contains the final value. The value is returned and the look up process is terminated.

In this example, two memory accesses were used to do the lookup, one in trie level 110 and the other in trie level 120. In practice, real tables contain many more nodes and levels than shown in this example. For instance, a typical forwarding table, in accordance with a preferred embodiment of the present invention, with 100,000 entries might contain 6 levels and 200,000 nodes.

In the preferred embodiment, the size of each entry within the nodes is fixed at 8 bytes and is independent

of the size of the key. This enables the internal memory width to be set to 8 bytes so that it is useful as ordinary memory when used in a bypass mode. A typical format of a node entry may be as shown in Table I.

field	bits	usage
bcnt	4	number of key bits used to index next node
scnt	4	number of key bits to skip
sbits	15	value to check against key when skipping
bindx	22	location of next node

TABLE I

If, for example, all the bits of bcnt is set to one, the remaining bits in the entry represent a value (either an actual value or the special value for lookup failure). This means that values can contain up to 60 bits. It also means that $1 \leq \text{bcnt} \leq 14$, so the maximum node size is 2^{14} entries. If any one of the bits of bcnt is not set to one, the entry represents a pointer to another node.

The depth of a trie depends primarily on the number of entries in the table and the distribution of the keys. For a given table size, if the keys tend to vary mostly in their most significant bits, the depth of the trie will be smaller than if they tend to vary mostly in their least significant bits. A branch of the trie terminates in a value entry when the bits that were used to reach that entry determine a unique key. That is to say, when there does not exist two different keys with the same leading bits.

The nodes of a trie can contain many empty entries. Empty entries occur when not all possible values of the bit field used to index a node exist in the keys that are associated with that node. For such routing tables about half the nodes are empty. Since, in the preferred embodiment, the size of a node entry is 8 bytes, such tables will consume about 16 bytes of memory per table entry.

Each trie entry in the look up table, according to the

10074022-22042001

embodiment of the present invention, includes a skip value field. During the lookup process, the skip value field is compared to the skipped key bits, and a lookup failure is signalled if they do not match.

5

10 The table lookup engine comprises at least one interface unit. The interface unit comprises an initiator and target interfaces to connect to a bus system of a processing system. The initiator comprises
15 a control and status interface for initialization, configuration and statistics collection, which is in the peripheral virtual component interface (PVCi) address space. There is a lookup interface for receiving keys and sending results of lookups, which is
20 in the advanced virtual component interface (AVCI) address space. There is a third memory interface that makes the internal memory of the table lookup engine available as ordinary memory, which is in the AVCI address space. All these interface units can be used
25 concurrently. It is possible to make use of the memory interface while the table lookup engine is busy doing lookups. Indeed, this is how the tables in the table lookup engine are updated without disrupting lookups in progress. The table lookup engine can be configured to use external (to the block) memory which can be accessed by the bus, in addition to or instead of its internal memory.

30 There are several internal registers that can be read or written. The control interface provides the following functions. Note that the key and value sizes are not configurable via this interface. The application that generates the tables determines how many key bits will actually be used. In the preferred
35 embodiment, the processing system supports key sizes of 32, 64 or 128 bits, but internally the table lookup engine expands shorter keys to 128 bits, by appending extra lower-significance bits. The table lookup engine always returns 64 bit values, but it is up to the
40 application how many of these bits to use.

10074033-0044001

Field	Bits	Reset Value	Function
Reset	0	0	When set (reset = 1), perform a complete reset of the table lookup engine blocks
Enable	1	0	When set (Enable = 1), enable the table lookup engine
CountersOn	2	0	When set (CountersOn = 1), enable updating of the counters

TABLE II

Field	Bits	Reset Value	Function
ResetStatus	0	0	Indicates whether the table lookup engine is busy resetting or not
EnableStatus	1	0	Indicates the enable state of the table lookup engine
CountersOnStatus	2	0	Indicates whether updating the counters mode is enabled

TABLE III

5

Field	Bits	Reset Value	Function
IntMemoryStart	31:0	note	Start location of internal memory in bytes
IntMemorySize	31:0	note	Size of internal memory in bytes

TABLE IV

10 Note: After reset, these registers contain the start and size of the entire internal memory. The application can change these if it wishes to reserve some portion of the memory for non-table lookup engine purposes.

Field	Bits	Reset Value	Function
ExtMemoryStart	31:0	0	Start location of internal memory in bytes
ExtMemorySize	31:0	0	Size of internal memory in bytes

TABLE V

Field	Bits	Reset Value	Function
NumLookups	31:0	0	Number of lookups
NumIntMemReads	31:0	0	Number of internal memory reads
NumExtMemReads	31:0	0	Number of external memory reads
NumIntBankReadsN	31:0	0	Number of reads by internal memory bank (N registers, where N is number of banks)
NumExtBankReadsM	31:0	0	Number of reads by external memory bank (M registers, where M is number of banks)

TABLE VI

- 5 The table lookup engine internal memory according to the embodiment of the present invention is organized as two equal size, independent banks. The size of these banks is a synthesis parameter. They are organized as a configurable number of entries with a width of 8 bytes.
- 10 The maximum number of entries that can be configured for a bank is 131072, which implies a maximum total memory size of 2 megabytes. Clients can use the table lookup engine internal memory in the same way as ordinary memory, bypassing the lookup state machines.
- 15 The address for a memory access selects one or more entries (depending on the details of the bus transaction) for reading or writing.

- 20 The protocol for a lookup is an AVCI write transaction to address TLEKeyAddr. Multiple keys can be submitted

for lookup in a single write transaction. The table lookup engine responds by sending back an AVCI read response to the source interface containing the values.

- 5 The table lookup engine has a key input FIFO with at least 128 slots, so it can accept at least that many keys without blocking the bus.
- 10 Lookups that succeed return the value stored in the table. Lookups that fail (the key is not in the table) return a special "missing value" containing a bit pattern specified by the user. It is feasible to construct the tables in such a way that a lookup failure returns additional information, for example, 15 the number of bits of the key that do match in the table. This assists the processing system in evaluating the cause of the failure.
- 20 The table lookup engine does not internally support longest prefix matching, but that effect can still be achieved by constructing the tables in the proper way. The idea is to split the overlapping address ranges into disjoint pieces.
- 25 Lookup values may not necessarily be returned in the order of the keys. The transaction tagging mechanism of AVCI is used to assist client blocks in coping with ordering changes.
- 30 Multiple client blocks can submit lookup requests simultaneously. If this causes the input FIFO to fill up, the bus lane between the requestor block and the table lookup engine will block temporarily. The table lookup engine keeps track internally of the source port 35 of the requestor for each lookup, so the result values will be sent to the correct place. This may to return the result to the requestor or elsewhere.
- 40 The contents of the memory being used by the table lookup engine can be updated while lookups are in progress. The actual updates are done via the memory interface. A software protocol is adopted to guarantee table consistency.

1007462-22042001

The table lookup engine 200, as shown in figure 2, comprises an input FIFO buffer 202 connected to the input of a distributor 204. the output of the distributor is connected in parallel to a plurality of lookup state machines 206a, 206b, 206c, 206d. Each lookup state machine 206a, 206b, 206c, 206d has access to a storage means. The storage means comprises a memory arbiter 208 and a plurality of parallel independent memory banks 212a, 212b. Each lookup state machine 206a, 206b, 206c, 206d is connected to the input of a collector 210. The output of the collector 210 is connected to an output FIFO buffer 214.

The table lookup engine uses a number of lookup state machines (LSM) 206a, 206b, 206c, 206d operating concurrently to perform lookups. Incoming keys from the bus are held in an input FIFO 202. These are distributed to the lookup state machines 206a, 206b, 206c, 206d by a distributor block 204. Values coming from the state machines are merged by a collector block 210 and fed to an output FIFO 214. From here the values are sent out on the bus to the requestor.

The entries of the input FIFO 202 each contain a key, a tag and a source port identifier. This FIFO 202 has at least 128 slots, so two clients can each send 64 keys concurrently without blocking the bus lane. Even if the FIFO 202 fills, the bus will only block momentarily.

The distributor block 204 watches the lookup state machines 206a, 206b, 206c, 206d and sends a key to any one that is available to do a new lookup. A priority encoder may be used to choose the first ready state machine.

The lookup state machines 206a, 206b, 206c, 206d do the lookup using a fixed algorithm. They treat all keys as 128 bits and all values as 60 bits internally. These sizes were chosen somewhat arbitrarily. It would be possible to extend the maximum key size to 256 bits. The main impact on the table lookup engine would be an increase in the size of the input FIFO 202 and LSMs 206a, 206b, 206c, 206d. It would be possible to increase the maximum size of the result. The main impact would be that trie entries would be larger than

8 bytes, increasing the overall table lookup engine memory required for a given size table. Shorter keys are easily extended by adding zero-valued least significant bits. Memory read requests are sent to the memory arbiter block 208. The number of memory requests needed to satisfy a given lookup is variable, which is why the table lookup engine may return out-of-order results.

- 10 The collector block 210 serialises values from the lookup state machines 206a, 206b, 206c, 206d into the output FIFO 214. A priority encoder may be used to take the first available value.
- 15 The memory arbiter block 208 forwards memory read requests from the state machines 206a, 206b, 206c, 206d to the appropriate memory block 212a, 212b. This might be to an internal memory bank or an external memory accessed via the bus. The table lookup engine has an
- 20 FBI initiator block for performing external memory reads. If the block using the table lookup engine and the external memory are on the same side of the table lookup engine, there will be bus contention. Avoiding this requires a bus layout constraint: the table lookup
- 25 engine must sit between the main processing units and the external memory, and the table lookup engine initiator interface must be closest to the memory target interface. Whether or not a memory read request goes to off-chip memory is determined by the external
- 30 memory configuration registers.

The output FIFO 214 contains result values waiting to be sent to the requestor block. Each slot holds a value, a tag and a port identifier. If the table lookup

35 engine received more than one concurrent batch of keys from different blocks, the results are intermingled in this FIFO 214. The results are sent to the correct clients in the order they enter the output FIFO 214, and it is up to the clients to use the tag to properly

40 associate keys and values.

Lookup algorithm

```

ValueTpe lookup(const lcsnode* trie, KeyType key)
{
    // toplgwd is size of level 0 node
45    int idx = key.topbits(toplgwd);
    key = key<<toplwd;

```

```

        lcsnode nd = trie[idx]; // current entry
        while (nd.bcnt != 15) {
            unsigned int cksbits =
key.topbits(nd.scnt);
5           key = key<<nd.scnt;
            if (cksbits != nd.sbits) return
missingValue;
            int nidx = key.topbits(nd.bcnt);
            key = key<<nd.bcnt;
10          idx = nd.bindx+nidx;
            nd = trie[idx];
        }
        return
concatenate(nd.scnt,nd.sbits,nd.bindx);
15    }

```

20 The table lookup engine according to the embodiment of
the present invention can achieve a peak performance of
about 300 million lookups/second. This level of
performance is based on the table lookup engine
internal memory system being able to sustain a memory
cycle rate of 800 million reads/second. This is
25 achieved by using two banks of memory operating at 400
million reads/second with pipelining reads. The latency
of the internal memory system needs to be of the order
of 4-8 cycles. The number of state machines is chosen
to saturate the memory interface. That is to say, there
30 are enough state machines so that one of them is doing
a memory access on nearly every cycle, for example 24
LSMs. Higher memory latencies can be tolerated by
increasing the number of lookup state machines, but the
practical limit is about 32 state machines.

35 The table lookup engine state machine lookup algorithm
is fixed and fairly simple, to attain performance. The
way that the table lookup engine achieves great
flexibility in applications is in the software that
constructs the LC-trie data structure. With this
40 flexibility comes a cost, of course. It is expensive to
generate the trie structure. The idea for using the
table lookup engine is that some general purpose
processor - for example in the control plane -
preconstructs the trie data and places it in memory
45 that is accessible by the bus, perhaps an external SRAM

block. An onboard embedded processing unit is notified that a table update is ready and it does the actual update in the table lookup engine memory. The table lookup engine state machines consider the memory it uses to be big-endian. When constructing trie structures the correct type of endianness needs to be employed. In this way the table lookup engine can provide longest prefix matching. When constructing the trie from the routing table, overlapping ranges can be identified and split. This preprocessing step is not very expensive and does not significantly increase the trie size for typical routing tables. It also allows multiple concurrent tables to exist. This is achieved by prepending a small table identifier to the key. With eight tables, this would require three bits per key.

The table lookup engine according to the present invention can return the number of matching bits. The lookup engine returns whatever bits it finds in the last trie entry it fetched. Further, on a lookup failure that entry is uniquely determined by the lookup algorithm; it is the entry that would have contained the value had the missing key been present. The program that generates the trie structure could fill in all empty trie entries with the number of matching bits required to reach that trie entry. These return values could be flagged some way to distinguish them from lookup table hits by the generator program. Then the table lookup engine would return the number of matching bits on a lookup failure.

The table lookup engine according to the present invention also enables concurrent lookups and updates. One way to achieve this would be to have two versions of the table in table lookup engine memory simultaneously, and switch between them with a single write to a table lookup engine configuration register. Then lookups in progress will find either a value from the old version of the table or a value from the new version of the table. The embedded processing unit achieves this by first placing the new level 1--n nodes in the table lookup engine memory, then overwriting the level 0 node entry that points to the new nodes.

The table lookup engine according to the present invention also allows very large results to be produced. If a value for a given key needs to be more than 60 bits, an auxiliary table can be placed in the

10074022-034403

table lookup engine memory - actually any available memory - and an index into the auxiliary table placed in the table lookup engine value. The auxiliary table would then be read using normal memory indexing. This
5 is purely a software solution, and has no implications to the table lookup engine internal operation.

Although a preferred embodiment of the method and apparatus of the present invention has been illustrated
10 in the accompanying drawings and described in the forgoing detailed description, it will be understood that the invention is not limited to the embodiment disclosed, but is capable of numerous variations, modifications without departing from the scope of the
15 invention as set out in the following claims.

204720" 22072001